CHOOSING CHARACTER RECOGNITION SOFTWARE TO SPEED UP INPUT OF PERSONIFIED DATA ON CONTRIBUTIONS TO THE PENSION FUND OF UKRAINE

Prepared by USAID/PADCO under Social Sector Restructuring Project

Kyiv 1999

CONTENTS

LIST OF ABBRIVATIONS

OCR (Optical Character Recognition): system for optical recognition of **printed** symbols.

ICR (Intelligent Character Recognition): system for recognition of symbols **written to a template separately of each other**.

OMR (Optical Mark Recognition): recognition of **marks** (such as crosses or form fields, e.g. squares or circles, marked with ticks, crosses, etc.).

IS: information system.

PCRS (Personified Contribution Record System): system for keeping record of contributions.

TWAIN: the most common computer/scanner interface standard.

AWB-E: the program system "Automated Work Bench of Employer."

MRF (Machine-Readable Form): paper forms filled out with printed or stylized symbols that may be automatically entered using optical scanners and the contents of which may be recognized using one of the automated recognition systems.

INTRODUCTION

In 1999, the Pension Fund of Ukraine (PFU) started introducing a personified contribution record system (PCRS) for insured individuals and corresponding accounting tools. It is expected that employers will submit data on contributions from their employees either on electronic media (if the enterprise has computers) or special paper forms. As it has become clear since early stages of implementation of the system in Lviv Oblast, only medium to large enterprises typically have computers, whereas these are virtually absent in small-scale enterprises, especially in rural areas. This means that such employers will report on the contributions only on the paper forms. Thus, operators of PFU's raion offices will bear the main burden of entering such data into PCRS.

The present report focuses on tools to input data from paper forms which would significantly streamline and speed up the processes of data input and handling.

1. TYPES OF INFORMATION SYSTEMS

All developed nations are currently introducing large-scale information systems (IS) allowing to accumulate information in databases, process it and provide to users in a proper form. Data processing and presentation procedures are usually completely automated as opposite to the process of information input (unless it is entered from magnetic media). From the point of view of the data update frequency, such systems may be divided into the following groups:

1) systems with rare changes in the contents of databases;
2) systems where the information is updated as the circumstances change (e.g. banking systems);
3) systems where the data is updated periodically, i.e. after a certain period is over (month, quarter, year, etc.).

The information input is not a crucial factor for the first-type systems. For those of the second type, the problem is solved using a regular team of operators. For the third group of systems, there is no need to employ permanent operators since these are needed only when the given period ends and then are no more necessary until the next period is over (though, the volumes of information to be entered may be quite large). PCRS may be cited as a typical example of IS belonging to the third category.

Earlier, the information input into IS databases was done either from pre-processed machine media or manually by operators from paper forms, using respective IS interface software. However, the prior preparation of the machine media was also done by operators, i.e. the information input in any case implied manual work, significant man-hour inputs, and inevitable errors while entering.

Thus, the urgent need to automate input of information from paper forms (primarily, figures and texts), recognition and entry to databases appeared long ago.

In most cases, operator has to enter information that is printed on a form, written into form fields, or presented as marks in such fields.

While in the past hardware and software did not allow to solve the problem of quality input and recognition of texts, lately there has been significant progress made in the development and application of industrial-grade (i.e. operating around the clock) optical scanners and automated text recognition systems. Such scanners come with a device for automated feed of pages from a package of documents and are capable of fast input of the documents' graphic into the memory of a computer.

In this report we consider only those systems which may be used to automate recognition and preparation for writing into a PCRS database of information entered from paper forms using optical scanners.

2. ANALYSIS OF EXISTING SYSTEMS FOR AUTOMATED TEXT RECOGNITION

2.1. Classification of automated text recognition systems

Currently used automated text recognition systems (ATRS) may be classified according to their recognition capabilities as follows:

- OCR (Optical Character Recognition): systems for recognizing printed symbols;
- ICR (Intelligent Character Recognition): systems for recognizing stylized symbols (stylized symbols are symbols written by hand separately of each other to a template, e.g. as on a postal envelope);
- OMR (Optical Mark Recognition): recognition of marks (marks mean crossed or ticked squares or circles in corresponding positions of a paper form).

In order to solve the text recognition problem for the purposes of PFU, one should be oriented towards ATRS already piloted and successfully used for similar tasks. Since the information on enterprises and personal data (such as wage, contribution amounts, period of employment, etc.) on insured individuals must be submitted to raion offices of PFU on special paper forms, systems allowing automated input and recognition of the contents of forms filled out with printed or stylized symbols are of greatest interest. Hereinafter, the forms that are filled out with printed or stylized symbols and may be automatically entered using optical scanners and the contents of which may be recognized using one of recognition systems will be referred to as machine-readable forms (MRF).

The quality of operation of any ATRS depends on the character recognition methods, based on artificial intelligence technology, most renowned of which include: structure-spot, multi-font, neural network, etc. These methods allow to adapt a system for recognition of alphabet symbols of any language and even any handwriting.

An analysis of existing ATRS with such abilities has shown that most common on the world market are presently the following systems:

- OmniPage produced by US-based Caere;
- Readiris produced by Belgium-based Softline;
- Recognita produced by Hungary-based OCR System;
- TextBridge Scan produced by US-based Xerox;
- Teleform produced by US-based Intelliscan;
- Presto produced by Russia-based ABBYY(Bit Software);
- FineReader produced by Russia-based ABBYY(Bit Software);
- CuneiForm produced by Russia-based Cognitive Technologies.

Also noteworthy is the work done by an L.M. Kasatkina-supervised team of the Cybernetics Institute of the National Academy of Sciences of Ukraine that has developed Ukrainian-oriented character recognition software based on neural networks of a rather high recognition level However, this system currently cannot compete with other ATRS since, at the present stage, it is more R & D than an industrial-grade software

product and does not provide for a complex solution of the problem of large-scale input and recognition of stylized symbols.

Far from all of the above ATRS are Cyrillic-compatible, and as for recognition of the Ukrainian alphabet, only **CuneiForm** of Cognitive Technologies and **FineReader Rukopis** of ABBYY (Bit Software) can provide for this. These ATRS may be attributed to the OCR/ICR class that completely meets the data input requirements for ATRS. An adaptation of any other ATRS for recognition of the Ukrainian alphabet is practically impossible without involvement of the manufacturer and will apparently require large costs.

Taking into consideration the above, only these two latter systems will be compared further in this report.

A survey of literature on ATRS has shown that these, to a certain extent equal, systems have been competing on the Russian market for more than three years. Together, they occupy not less than 95 percent of the market. Initially, the product of Cognitive Technologies gained the lead. Later, however, FineReader abruptly outran its competitor and is still leading among this class of software on the Russian-lingual market of former Soviet Union, as well as in some other countries.

Presented below is a summary of what has been found in the media, promotion and Internet, as well as on CDs provided by the companies' representatives concerning the two ATRS most popular in CIS. More detailed information is provided in the Annexes.

3. ATRS BASIC CHARACTERISTICS

3.1. CuneiForm

3.1.1. Some information on Cognitive Technologies

The company Cognitive Technologies Ltd. was established in 1993, being based on the Artificial Intelligence Laboratory of the Russian Academy of Sciences' Institute of System Analysis (former All-Union Research Institute of System Studies of the Academy of Sciences of the USSR). The laboratory was supervised by Prof. V.L. Arlazorov, Sc.D. in physics and mathematics.

3.1.2 CuneiForm characteristics

CuneiForm has the following characteristics:

- Recognizes forms printed typographically as well as with dot, jet, laser and other printers and filled out with printed text of stylized symbols (hand written).
- Recognizes marked questionnaire items.
- Is compatible with some 30 TWAIN scanner families.
- Automatically sets optimal scanning brightness.
- Supports streaming input mode in client/server architecture.
- Supports recognition of printed symbols (OCR mode) of the following languages: Russian, Ukrainian, English, mixed Russian-English, German, French, Spanish, Italian, Dutch, Danish, Swedish, Serbian and Croatian.
- Supports Russian and Ukrainian as the languages for handwriting recognition (ICR mode).
- Recognizes any fonts without training, hand-printed texts and stylized figures (the way of filling out should be inquired for in Cognitive Technologies).

- Automatically exports to any databases through DBF, CSV or ODBC. Automatically checks the correctness of recognition of form fields that must hold a value.
- Performs context check.
- Is rather easy to use and features a friendly interface.

The standard form recognition system is intended for streaming input of payment orders, tax returns, customs statements, insurance forms, and any other documents having a standard field layout.

The system has a well developed help system and documentation in Russian, a possibility for unlimited expansion, a warranty on all software and hardware supplied, as well as the supplier's technical support and training of the buyer's personnel.

The CuneiForm system is based on a client/server architecture enabling efficient real-time data access and a maximum efficiency in making use of hardware performance capabilities (see Fig. 1).

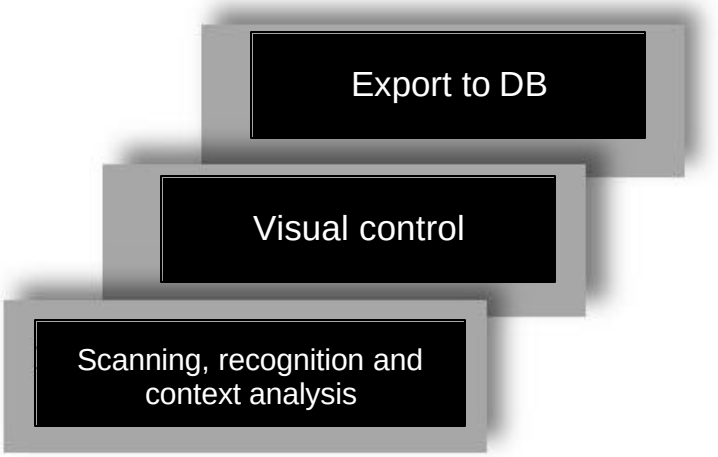


Fig. 1. Block scheme

The CuneiForm system consists of several subsystems. In the **Input Subsystem** (CognitiveForms), the Input Servers (CuneiForm 98NeSt) automatically manage recognition and its context check. The Visual Control subsystem (Cognitive FormEditor) works with the System Client (operator) and is responsible for visual check and editing of a document before export to database.

Cognitive Technologies has specially developed a special technology to increase the labor productivity, decrease the operator fatigue and control the correctness of document input. Subject to human control are only those data that do not meet recognition correctness requirements. Export to database is carried out by operator after input of certain number of forms. The database is customer-chosen among those existing.

The **Input Subsystem** provides for:

- Receiving of document image from a scanner or through a faxmodem;
- preliminary form processing (elimination of boxes, lines, shadings, etc. and isolation of standard data elements);
- recognition (document conversion from a graphic file into a text file);
- context analysis of recognition correctness;

- automated data adjustment (e.g. sum check, date check, numeric field check, field value legitimacy check with an external dictionary, etc.).

The forms are typographically made or printed out standard blanks filled out with a typewriter, printer or hand.

There are following basic requirements to forms:

1. The form of documents of the same type must strictly conform to an approved standard.
2. Field layout must be strictly fixed.
3. In case of the typographically made forms, the color of invariable areas (boxes, lines, shadings) must be light green or "weak gray" (depending on the scanner performance capabilities).
4. In case of computer-generated forms, rescaling is allowable depending on the type of printer with which these have been printed out.
5. Only muster copies must be processed.
6. Languages: Russian, Ukrainian, English, mixed Russian-English, or five other European languages.
7. Typeface is of no importance. Typographic forms are preferably to be filled out with a typewriter.

Check of correctness of information recognition is secured by a special way of presenting information on operator's monitor: an operator may simultaneously see the recognized text in the form field and the document input image received from a scanner or through a faxmodem. Fields with doubtful symbols or those automatically adjusted by the context control are highlighted as doubtful. This means that the operator does not need to review all form fields and compare these with the input document. Thus the operator fatigue is decreased and the correctness of information input in most important operations is increased.

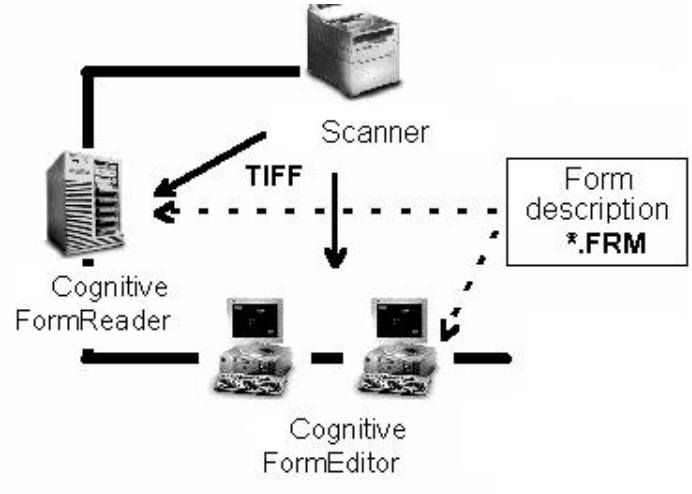


Fig. 2. CuneiForm process layout

In brief, the CuneiForm uses the following procedure:

1. Cognitive Technologies makes a description of one or several forms (a file with .frm extension) after your model.
2. Using any scanner, hard copies are entered into computer and stored as images (files with .tif extension).

3. Cognitive FormReader software is used to recognize standard forms meeting the Cognitive Technologies presentation requirements. After the recognition, the operator may check the correctness of recognition, adjust the data and store it in a format of required DB.

The company has provided the following results of an analysis of recognition performance:

- Input time for one A4 page containing 200 words: 3s to 30 s. When Pentium 100/32Mb RAM is used, this time depends on the scanning speed, the number of words to be recognized and the context processing complications.
- Percentage of correctly recognized printed or typed symbols in numeric fields: 99.6%.
- Percentage of correctly recognized hand written symbols in alphanumeric fields: 99.5%. In case of forms filled out with hand, the quality to a great extent depends on the context.

The PC-type hardware requirements:

1. Cognitive FormReader, Scan Station:
   - ISA, SCSI interface;
   - IBM-compatible i586-100 or higher;
   - at least 16Mb RAM (32Mb recommended);
   - MS Windows 95, NT 3.51 or later versions, Novell 3.12 or later versions;
   - 20Mb HD space available for software installation plus 1Gb to 4Gb for streaming operation.
2. Cognitive FormEditor:
   - i486 or higher;
   - at least 8Mb RAM (12Mb recommended);
   - MS Windows 95, NT 3.51, Novell 3.12 or later versions;
   - 500Mb HD space.

From the point of view of general system reliability it is desirable to provide for maximum equipment unification.

Virtually all basic equipment necessary to run CuneiForm is manufactured by the companies Hewlett-Packard, Kodak, Bank Tec and Bell+Howell:

- file server: HP NetServer;
- magneto-optical disk drive: HP SureStore;
- workstations: HP Vectra;
- any networking equipment.

All Hewlett-Packard products are warranty-covered from one to three years free of charge. Maintenance is provided by the HP Service Center in Moscow, Russia. Technical support for the entire line of equipment is provided by vendor company.

Kodak equipment is covered with a free warranty of one year. Further support is available through purchase of a service contract for a minimum of one additional year. Maintenance and technical support are provided for the entire product line by vendor company and the manufacturer directly.

Cognitive Technologies provides for a full complex of services including installation, necessary hardware follow-up, customer training, technical support, and turnkey system adaptation for the customer's specific task. Customers may not adapt the system for recognition of required forms on his/her own.

Any project implementation include the following phases:

- feasibility study;
- pilot project;
- full-scale design.

Examples of use of CuneiForm:

- mass media (input of questionnaires);
- *Hard&Soft* magazine;
- *ComputerPress* magazine;
- *Moskovskiy Komsomolets* newspaper;
- State Tax Inspection in the Bashkortostan Republic;
- two projects on input of tax card registers of legal entities (OCR, printed);
- a project on input of individuals' tax declarations (hand-printed);
- a system for large-scale processing of insured persons' forms and data on employment period and income for the Pension Fund of Russian (within Moscow).

Use of CuneiForm is exemplified below in the variants of application of an automated payment order forms (POF) input system.

- The typical variant is to supply basic software modules intended for automated POF input and, probably, respective services, such as technical support, training, consultations, etc. In this case, modules setup as well as the development and setup of a streaming input process will be done by the customer.
- The project variant of supply of an automated POF input system means that all the works on setup and adaptation of the software modules and streaming input process for the customer requirements will be done by Cognitive Technologies.
- The works on the development and setup of a streaming input process include (but are not limited with):
  - development of a flow dispatching module for documents being processed for recognition and editing stations;
  - development of a statistical module for collection/analysis of information on edited forms (questionnaires). The statistical files are formed in the process of operation of recognition and editing programs. Each entry in a statistical file contains information on the POF pack, handling process, process start and completion time, its results, operator, etc. An individual file is created for each workbench. Taken into account is the specificity of input technology existing with the given customer (user interfaces, connected dictionaries, dictionary actualization technique, processing allocation, hardware type, etc.).

The streaming input process may be outlined as follows:

- Preparation of POF packs for scanning (unstitching packs, flattening forms and loading these into input stacker).
- Automatic assignment of identification numbers for POF packs in streaming mode.
- Start of scan mode for a POF pack, visual control over the scan process displayed on monitor.
- Automatic save of scanned POF images in a server folder.
- Automatic allocation of the scanned POF images flow among recognition and editing stations.
- Recognition of scanned POF packs in streaming mode without visual control over the form recognition process.
- Sequential visual control and editing of field contents in a selected POF pack. Only doubtful (highlighted) fields and those user-preset will be checked, thus saving time.
- In case any erroneous POF are detected, the pack will be marked as "erroneous" and placed in a special folder for further processing. When its editing is completed (i.e. all errors are corrected) this pack will be converted into a required data structure.
- The file created upon the conversion will be saved in a special folder for final check by a special program. If the check results are positive, the pack will be considered suitable for loading into the DB buffer.

3.2. FineReader

3.2.1. Some information on ABBYY

The company ABBYY was established in 1989. With its headquarters in Russia, the company has subsidiaries in the United States (ABBY USA) and Ukraine (ABBY Ukraine) as well as exclusive agencies in more than 20 countries, including Germany, France and Australia. The Russian Computer Union has officially named it the number one software manufacturer in Russia.

To recognize symbols, so-called structure-spot standards are used. This method, called font transformation, was proposed in 1992 by David Yan, Konstantin Anisimovich and Pavel Senatorov. The team of scientists of the Research Center for Electronic and Computer Engineering that in 1977 through 1986 had worked on recognition of stylized inscriptions on drawings developed the most important approaches to solution of such problems.

3.2.2 FineReader characteristics

Since 1993, when FineReader started being developed, four generations of the ATRS has been created. The most recent one includes FineReader 4.0 Standard (OCR), FineReader 4.0 Professional (OCR) and FineReader 4.0 Rukopis ["Manuscript"] (ICR) that feature as follows:

- Recognize forms printed typographically as well as with dot, jet, laser and other printers and filled out with text printed or hand written.
- Recognize marked questionnaire items.
- Support scanners with TWAIN interface.
- Support streaming document input and processing mode.
- Support recognition of printed symbols (OCR mode) of 40 languages.
- Support recognition of stylized symbols of the Russian and Ukrainian languages (ICR mode).

- Automatically export recognized information to databases having a DBF, CSV or ODBC structure. Automatically check the correctness of recognition of form fields that must hold a value.
- Carry out context control.

The system is also quite easily used and has a friendly interface. The system operation consists of two main phases: first it analyses the image acquired from a scanner (defining areas of recognition, tables and pictures, isolation of lines and individual symbols in the text), and then recognizes each symbol per se. The task-oriented search allows recognition of broken and distorted images, making the system resistant to possible page defects and nonstandard symbol outlines. Since self-learning is a preset algorithm feature, it is easily adapted for different fonts, still keeping quite a high recognition probability of more than 99.7 percent.

A special know-how is used for automatic input and automated recognition of the contents of forms filled out with stylized symbols hand-written by different individuals (ICR). It covers the entire process: from scanning hard copies to writing checked and corrected recognition results directly into a database.

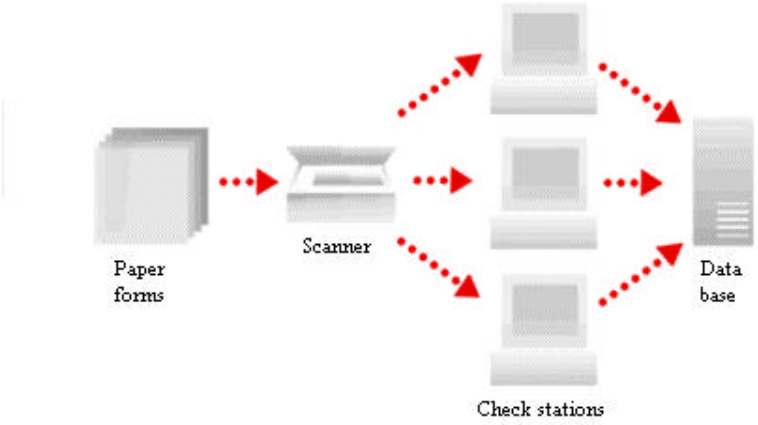


Fig. 3. MRF recognition system based on FineReader Rukopis

Any MRF developed to comply with the ABBYY's "MRF Design Requirements" may be an input document for the system. The paper form contents written in database will be the input information.

The system operation cycle goes through five phases:

1. Document scanning.
2. Entering the form image into a recognition package.
3. Document recognition and automatic control over the recognition results.
4. Visual control and correction of data rejected by the automatic control.
5. Transfer of the recognition results to database.

The above phases are described below with more details.

3.2.2.1. Document scanning

MRF are scanned at one or several scan stations using high-speed industrial-grade scanners (e.g. scanners of BankTec).

Operator loads a pack of documents into the scanner and starts scan function. When processing of the pack is over, the scanner will stop and the display will prompt the operator to load the next pack and continue scanning.

The document images thus entered are placed in an image array (a unique number is assigned to each image).

3.2.2.2. Entering the form image into a recognition package

Concurrently with the document scanning process, the FineReader Rukopis input module is continuously interrogating the image array for any new images. Once such image is detected, it will be immediately automatically moved to a package for further recognition. If no new image, the input module will switch over to standby mode, awaiting for a new image.

3.2.2.3. Document recognition and automatic control over the recognition results

After the image has been moved to the recognition package it will be processed by the FineReader Rukopis recognition module that is continuously interrogating the package, looking for unrecognized images. Once such image is detected, the recognition module will start processing it. If no unrecognized image in the package, the recognition module will switch over to standby mode, awaiting for arrival of a new image in the package. All actions are performed by the system automatically, without operator involvement. The recognition procedure takes three stages:

1. *Optical text recognition*. The form fields layout is automatically identified and the fields are recognized using the algorithms of optical text recognition. The font transformation technique is used at this stage.
2. *Spell-check*. The orthography of recognized text is checked. Wrong or dubious symbols and words are marked with another color.
3. *Automatic check of the recognized text for conformity with control rules.* At this stage, the system automatically checks the recognized text for its conformity with the control rules that were described when setting FineReader Rukopis up for the structure of specific MRF.

   The control rules below may be applied simultaneously:

   - To a dictionary or a database. Checking the field value against a dictionary or a database, For example, name of locality may be checked against a dictionary of localities.
   - To a template. Checking the field value against a regular expression (PERL). For example, passport number and series must conform to certain template.
   - Cross-validation. Carrying out cross-validation of several database fields. For example, if the name (family, given, patronymic) and the identification number of a client are entered, it will be possible to check if these match.
   - Sum in numbers/sum in words. Identity of the sums written in numbers and in words (for the Russian and Ukrainian languages).
   - Check sum. The figure in the “Sum” field must match the sum of the “Total” fields.

3.2.2.4. Visual control and correction of data rejected by the automatic control

The documents in recognition of which there emerged any errors or unconformity to the control rules are marked as erroneous. The operator is made informed of the error place in each rejected document.

The operator checks the rejected documents on the computer monitor. Doing this, he can simultaneously see the results of recognition of all the form fields and the form image. The current symbol is constantly selected on the image with a color rectangle. This allows correction of the recognition errors not referring to the hard copy.

3.2.2.5. Transfer of the recognition results to database

Having the recognition results checked, the package of recognized forms will be automatically transferred to the FineReader Rukopis export module. Entry to the database is made by the operator after having checked each sequential batch of documents.

In addition, the FineReader Rukopis system is able to input several types of MRF within the same stream of documents. The program automatically selects the right template for each document. An individual table in the database may be linked to each form type.

As for the system versatility, it should be noted that all the workstations involved in the process of document input and handling may carry out the functions of scanning, recognition and visual control. Depending on changes in the intensity of unrecognized or recognized document flow, these functions may be redistributed among the process participants.

4. TESTING

Described below are the results of testing ATRS CuneiForm and FineReader to compare their recognition abilities for printed symbols (OCR) published in *Komputernoye Obozreniye* ["Computer Review"] #13, 1998, and the results of testing the two by PADCO to compare their stylized symbol recognition performances (ICR).

4.1. OCR testing materials and methods

The tests were carried out using a computer HP Vectra VE (Pentium 166 MMX, 32Mb SDRAM, MS Windows 95 OSR2) and a scanner HP ScanJet 4c. The OCR built-in scanner support was used. The TWAIN interaction mode was attached only in FineReader 4.0. The text documents were scanned at a resolution of 300 dpi. The first test – for minimum symbol size – was carried out using three documents containing a text at different font sizes of 12, 10, 8, 6 and 4 points printed out with a laser (HP LaserJet 5L, 600 dpi), a jet (Epson Stylus Color Pro, 720 dpi) and a dot (Epson FX-1170, 120×144 dpi) printer. The second test – for low-quality text – also included three documents: too light a fax of bad quality (an uneven text with some faint symbols), a newspaper article (a multicolumn text, low-quality paper and types) and a text of a complex structure made in small print on stamp paper (a complex, polychromatic background). The third test was intended for recognition of a complex table created in MS Word 97. In all the tests, the number of errors per 1,000 symbols was counted. The tests were repeated thrice. In case of the table, the results were evaluated qualitatively.

The versions FineReader 3.0, Fine Reader 4.0 and CuneiForm 98 were used for testing.

4.2. ICR testing materials and methods

The abilities of the two ATRS were tested in PADCO using a computer Cel 300a/440LX/4Mb AGP/4.3Gb/36x CD, MS Windows 98 and a scanner Mustek ScanExpress 60000SP in the TWAIN mode. The test documents were scanned at a resolution of 300 dpi.

The first test was carried out using an INDANI (individual data) form that is presently used in PCRS for reporting on wages, contributions to PFU and employment period of the insured.

The second test was carried out on a form with the same contents but of somewhat different appearance (adapted for the requirements of the recognition system).

The third test was carried out on specimens of the forms used in the Pension Fund of Russia that had been provided by ABBYY.

The fourth test was carried out on specimens of the questionnaire forms from a demo version of Cognitive Technologies.

The versions FineReader 4.0 Rukopis and CuneiForm 98 were used for testing. All products used in testing were provided by official representatives of ABBYY and Cognitive Technologies in Ukraine.

5. TESTING RESULTS

5.1. Testing results from literature

The data in Tables 1 and 2 was published in the *Komputernoye Obozreniye* magazine and shows the percentage of erroneously recognized symbols in testing.

Table 1. Recognition of documents printed out with different printers (% of errors)

Printer	Laser Jet 5L					Epson Stylus Color Pro					Epson FX-1170				
Font size, pt	12	10	8	6	4	12	10	6	6	4	12	10	8	6	4
FineReader 3.0	0.00	0.00	0.00	0.10	0.20	0.0	0.0	0.0	0.1	11.5	0.3	0.3	0.9	1.7	>50
FineReader 4.0	0.00	0.00	0.00	0.00	0.00	0.0	0.1	0.2	0.7	1.3	0.2	0.2	0.4	1.3	>50
CuneiForm 98	0.40	0.60	1.00	1.00	3.60	1.6	1.6	3.5	10.2	34.3	4.8	27.6	49.1	>50	>50

Table 2. Recognition of low-quality texts (% of errors)

	Fax	Newspaper	Text against background
FineReader 3.0	1.4	1.9	4.9
FineReader 4.0	1.9	3.7	28.7
CuneiForm 98	>50	10.3	>50

As demonstrated by the above tables, the ABBYY products showed their decisive superiority in the first test. The laser printer text brought about no problem for any of the two FineReader versions, whereas CuneiForm made as many as 3.6 percent of mistakes when recognizing 4pt font. FineReader 4.0 excellently recognized the text printed out with the jet printer with a minor percentage of errors only with 4pt font. Version 3.0 of the program missed more often with this font: 11.5 percent of errors.

CuneiForm showed a geometric progression in worsening the recognition quality with the decrease of font size: it failed to recognize about 34.3 percent of characters in the 4pt-font text. The last document – made with the dot printer – turned most illustrative in this test. None of the OCR programs was able to cope with 4pt font: more than 50 percent of errors with all three. CuneiForm had an acceptable recognized accuracy only with the text of the maximum font size, whereas both FineReader products made just few mistakes within the range of font sizes 6pt to 12pt.

Processing the bad-quality fax, CuneiForm showed too bad a result. Whereas FineReader stumbled on 1.4 percent to 1.9 percent of the characters, the Cognitive product obviously failed (more than 50 percent of errors).

The paper article brought better results for CuneiForm than the fax did, though the error rate was also rather high: 10.3 percent.

It should be noted that FineReader 3.0 heads the list in the speed of recognition. Version 4.0 of the program is second best, loosing not too much. CuneiForm fell far behind. Besides, it all the time gave a message of having not enough memory for completion of the operation.

When it came to processing the test table, only the latest version of FineReader showed excellent results: the table completely preserved its appearance. FineReader 3.0 also recognized the table but its form took a standard rectangular shape. CuneiForm failed to pass this test at all, in spite of setting the Table option and subsequent manual marking of blocks.

The background test turned out difficult for FineReader 3.0 and the more so for CuneiForm that flunked one more test as a matter of fact. The fourth version of FineReader was quite successful at this stage, primarily owing to image scavenging. Though, all punctuation marks were cleared out along with the "garbage."

5.2. PADCO testing results

The results of recognition of the three forms mentioned under 4.2 are shown in Table 3.

Table 3. Recognition of stylized symbols in the forms (% of errors)

	FineReader	CuneiForm
1. INDANI form	>90	>90
2.Modified INDANI form	>50	>70
3.Russian PF form (FineReader)	5	50
4.Questionnaire form (CuneiForm)	5	25

The first test (recognition of the INDANI contents) showed it straightway that the current form is unfit for automated recognition.

In the second test, almost the same form was used, however, the size of its text blocks intended for recognition had been enlarged. The results of this test may be reckoned unacceptable only for CuneiForm (more than 50 percent of errors).

The third test showed very good results for FineReader, at only 5 percent of errors, due to thorough preparation of the form template and blocks description, but only provided that the proprietary blanks and settings as specified in the system's demo versions were used. To a certain extent, this may be creditable to the very competent support provided by ABBYY Ukraine, specifically Mr. M. Tkachenko who supervised the setup. In our opinion, all our attempts to get similar results with CuneiForm have failed because only experts of Cognitive Technologies can adapt forms (see. Annex 5). In this connection, we carried out the fourth, additional test, where forms developed by Cognitive Technologies were used, to demonstrate the capacities of CuneiForm.

As was shown in the fourth test, CuneiForm had a very high level of recognition (1 percent of errors) for own demo files, but when it came to new forms – printed out, scanned and additionally filled out for recognition – the level dropped (25 percent of errors). As for FineReader, it again demonstrated its steadily good results (only 5 percent of errors) that could have been even higher had the dictionaries been prepared more thoroughly.

It should be noted that fine-tuning to a form template in any ATRS, whether FineReader or CuneiForm, requires large labor inputs (about 4 to 5 man-months per form for experts of a company).

The makers of CuneiForm believe that users ought not to be involved in such tuning in principle (see. Annex 5). ABBYY also offers its services to adapt the system for recognition of specific forms, however, it allows involvement of users in the process (see Annex 6). Obviously, such a job would be done to tune a recognition system to a template of a very mass form (tens of hundreds thousands of copies). Such labor inputs had not been budgeted for this experiment and therefore, when making preparations for testing, some works were done not thoroughly, e.g. description of dictionaries of the objects admissible in form fields as well as valid or invalid symbols. As for CuneiForm, such works cannot be performed by a tester

at all since, as mentioned above, the procedures for description of these dictionaries are not described in the user manual and can be only carried out by experts of Cognitive Technologies under a software supply contract. Hence, the above empirical data should in no way be taken as a "final diagnosis" for one or another of the examined ATRS. If such a work is necessary further on, it may include refining the contents, size and other form characteristics, as well as coming to terms with ATRS manufacturers.

6. CONCLUSIONS

1. CuneiForm 98 and FineReader 4.0 Rukopis may be considered most suitable for the purposes of the Pension Fund of Ukraine among text recognition systems of the OCR/ICR type.
2. Testing has showed that use of CuneiForm 98 (i.e. the version used in testing without adaptation for font) will be justified only for recognition of good-quality texts. Its value being the same as that of FineReader, this will hardly make choice of CuneiForm justified. CuneiForm may be regarded as a competitor of FineReader only in view of its free distribution along with scanners. However, this can be hardly taken as a telling argument for PFU, since it would have anyway to conclude a contract with the company to adapt for necessary forms.
3. According to the results of a literature analysis (articles, reviews in specialized journals, etc.) and Internet search, as well as reports from users and the results of testing, it is possible to come to a conclusion that FineReader 4.0 Rukopis is currently most suitable for input and recognition of the forms of reporting on wages, contributions to PFU and employment period.
4. The important advantages of FineReader also include the presence in Ukraine of a subsidiary of the manufacturer (ABBYY Ukraine) employing skilled personnel. This allows further tuning of the system to required forms independently, without the software creator company, making use of advice when necessary but not entering into agreement on entire work.

ANNEX 1. PROJECT REALIZATION EXPERIENCE WITH THE SYSTEM FOR INDUSTRIAL FORM INPUT ABBYY FINEREADER RUKOPIS

The users include: the RF Tax Police Federal Service, the RF State Tax Service, the Pension Fund of Russia, the Central Bank of Russia, the Savings Bank of Russia, and others. Overall more than 100,000 legal users in 15 countries.

Presently, *FineReader Rukopis*-based systems are used in the *State Tax Inspection* in Moscow where two powerful industrial systems for input of tax documents (tax declarations, reporting) at a capacity of 10 declarations per minute have been installed. A similar system is being installed in St. Petersburg. There are less powerful systems in the tax inspections of the oblasts of Moscow, Tula, Nizhny Novgorod, Rostov and Volgograd, as well as of the Yamalo-Nenetsk Autonomous District. The capacity of a system with one operator is about 60 thirteen-paged declarations per hour, and with two operator, 120 declarations per hour. These figures have been officially proved in numerous tests, e.g. in the Tula oblast inspection. For certificates of incomes, the rated capacity is about 360 certificates per hour per one operator.

MRF blanks *Declaration of Incomes* for 1998 and 1999 have been developed for the RF State Tax Service. The blanks developed by the company are in full conformity to international standards for MRF.

The *Pension Fund of Russia* has already bought five industrial document input systems developed by the company. By now these have been installed in Moscow Oblast, St. Petersburg, Krasnodar Krai, Krasnoyarsk Krai and Novosibirsk Oblast. These regions are on the top of the regional automation list approved by the RF Pension Fund. The automated input system in Moscow Oblast has been industrially run for more than one year. The systems installed in other regions are at a stage of trial operation, except for Krasnodar where transition to full-fledged operation is taking place. According to research done by the Pension Fund, the recognition quality for the entire range of non-checked documents is on average 99.8 percent, and for the documents meeting fill-out requirements, 99.93 percent. After a check the rate of errors is below 0.001 percent. The RF Pension Fund has officially confirmed the data.

ABBYY has developed for the RF Pension Fund a number of MRF blanks, such as *Insured Person Form*, *Individual Data on Employment Period and Earnings*, and others. These blanks are used in all offices of the Pension Fund or are introduced in the process of automation of offices.

ABBY has developed the following forms and blanks that are currently used in Russia:

Organization	Form title	Number of copies
Pension Fund	Insured person form	150,000,000
Pension Fund	Individual data on employment period and earnings	200,000,000 annually
State Tax Service	Declaration of incomes	7,000,000 × 13 pages
State Tax Service	Certificate of incomes	20,000,000
Savings Bank	Form 187 AC (payment order)	10,000
Federal Center for testing high school leavers	Questionnaire	400,000 annually
Moscow Government's Small Enterprise Department	Application for entry in the Register of Small Entrepreneurship Entities of Moscow	300,000 to 600,000 annually

ABBYY has developed and put into operation the following large-scale process lines for document input that are currently used in Russia:

Organization	System title	Capacity
Federal Service of the Tax Police	Standard process line for input of special-purpose documents	up to 30,000 sheets/day
State Tax Service	FineReader Tax	up to 100,000 sheets/day
Pension Fund	FineReader for Pension Fund	100,000 sheets/day
National News Service	Process line for document input	20,000 sheets/day
National Registration Company	Process line for document input	1,500 bulletins/10 minutes
Federal Center for testing high school leavers	Process line for document input	400,000 annually
Moscow Government's Small Enterprise Department	Standard process line for input of documents	300,000 to 600,000 annually

ANNEX 2. ABBYY REQUIREMENTS TO MRF

1. Requirements to paper forms

MRF is an ordinary paper form (questionnaire, blank, etc.) made to meet the requirements specified herein.

As any other form, MRF is intended to fix a primary *subject information* from the *person filling it out*. However, unlike majority of ordinary forms, MRF enables input of the information into databases (DB) in an automated way using industrial scanners and software for optical character recognition (OCR).

MRF may accept:

1) being filled out with hand or printer/typewriter;

2) being filled out only with printer/typewriter.

The appearance of MRF in the first case differs from the other one by larger fields.

FineReader allows inputting the following types of subject information:

Description	Explanation	FineReader 4.0 Rukopis
Hand printed letters	Written with hand in print letters	Russian, Ukrainian, English
Hand written figures	Normal figures written with hand or figures written to a template, e.g. as on envelopes	Yes
Printed text	Typographically, with a dot printer, typewriter	Rus., Engl., Ger., Fr., Ukr., etc. (altogether 40 languages)
Items	Ticks	Yes
Groups of items	One mark of several	Yes

In addition, the system allows predefining subsets of symbols and words for each field. For example, when recognizing an address in the "City" field, the words from a list of cities may be recognized. The system also allows linkage between fields for automatic control against an external database, field masks, control sums, etc.

Thus:

1. All MRF of one type must be printed from one original (be identical when viewed in transmitted light). A tolerance for the linear dimensions of form elements is 0.15 percent (half a millimeter of the A4 page height).
2. MRF must have *benchmarks* for the OCR software to carry out adjustment after scanning (correcting misalignment, compensating linear and nonlinear scanning defects, matching templates). For the requirements to the benchmarks, see below.
3. Any *explanatory information* (any information and all graphical elements present on a blank form: headings, field captions, boundaries, pictures, benchmarks, etc.) must be executed in a way so that it does not interfere with the subject information. The explanatory information requirements are provided below.
4. It is recommended providing *abundant subject information* for important fields to enable control over correctness of form filling-out and its recognition by the system. The information excess is realized traditionally: sum in numbers/sum in words, control sum, client name/client account, etc.

Special markers, such as *black squares*, dividing lines or a static text (text of explanatory information) may be used as benchmarks.

The benchmarks should meet the following requirements:

1) dimensions 4×4 mm to 7×7 mm;
2) no rectangles other than squares are acceptable;
3) for forms of the same type, all squares must be of the same size;
4) the squares must be placed in pairs, strictly aligned horizontally;
5) the space between any square edge and the nearest object (text block, line, picture, etc.) must be not less than 3 mm.

2. Requirements to explanatory information

As mentioned above, *explanatory information* (any information and all graphical elements present on a blank form: headings, field captions, boundaries, pictures, benchmarks, etc.) must be executed in a way so that it does not interfere with the subject information. There may be two types of explanatory information (EI): (1) appearing on the image after scanning and (2) disappearing after scanning.

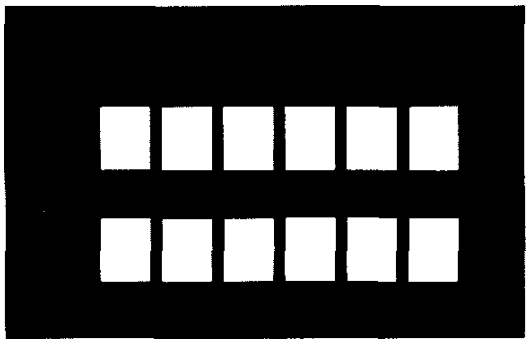
Orange form

Ideally, entire EI should disappear except for benchmarks. This is attained through having the EI printed in light orange and scanned in the Red-Lamp mode (available in majority of high-speed scanners).

Black-and-white form

In case of b/w pint it is necessary that *Input Field Flags (IFF)* for subject information disappear after scanning. To this end, the IFF are printed in either (1) light value of gray or (2) fine black dots.

Input Field Flags (IFF)



It is most advisable to form the fields for input of subject information as a series of white rectangles for letters against a gray background. The rectangle dimensions should be not less than 4×5 mm. The horizontal distance between the rectangles should be 1 mm or more, and the vertical distance between the rows of rectangles, not less than 4 mm.

The gray background should be either very light (5% of black) or made of individual dots not more than 0.1 mm at a spacing of about 1 mm.

3. Requirements to items and groups of items

The system allows recognition of items of different appearance: marks in squares, clusters or on a white field, and circled text, meeting the following requirements:

1) item dimensions from 2×2 mm to 10×10 mm.
2) the space between item edge and the nearest object (other item, text block, line, picture, etc.) not less than the size of the item.

FineReader does not impose any limitations on the item shape: square, rhombus, letter, etc. However, it is recommended to use square (☐) as an item and cross as a mark (☒).

4. Requirements to form filling-out

The forms ought to be filled out neatly, in capital letters, using a black-ink ballpoint or capillary pen. Indigo ink will also produce a good result. A felt-tip pen will result in a lower recognition quality because of merger of fine details. The worst recognized are pencils and pens with light-color ink. It is therefore strongly recommended placing on form the following text:

"WARNING! Please fill out with a BLACK or INDIGO-ink ballpoint or other pen, except for felt-tip pen, in CAPITAL PRINTED LETTERS."

5. Service notes

There may be envisaged a place for service notes on a form. This field is to be filled out by operator upon the receipt of documents. It may include fields like document running number, date of receipt, identity number of the person who received the document. Such notes may be made using a stamp, by hand or printed out, and then also recognized and entered into a database.

In case of a large number of MRF copies, it is recommended that the original be submitted to ABBYY/BIT Software for certification. This service if free. The original will be examined by experts of the company for machine readability on the basis of the dozens thousand forms input experience and available information on typical errors made by clients when filling these out.

In case the form has been certified, it may bear the following line:

"This form meets the FineReader MRF specifications."

ANNEX 3. ANALYTICAL CAPACITIES OF FINEREADER 4.0 RUKOPIS

Form package recognition with automatic template selection. FineReader 4.0 Rukopis allows processing an unsorted stock of forms of different appearance in the same package. To this end, up to 99 templates may be created in one package before scanning. When scanning such a package, the template for a given image will be selected automatically.

Alphabet limitation. Each form field may be made corresponding to its own text language that may be based on a word language or a combination of word languages (mixed languages as Russian-Ukrainian, etc.). There may be following word languages:

1) standard (Russian, Ukrainian, English, German, etc.);
2) obtained by limitation of standard symbols (e.g. figures);
3) based on standard languages but with a limited dictionary (e.g. the days of week, the months, years, family names, given names, etc.).

Automatic recognition correctness control. The form control allows minimizing the number of misrecognized forms that are difficult to detect by a simple review of the package. After the recognition of each form, the user can check what rules have not been observed and why. There are five preset types of rules (in addition, own rules may be automatically created as DLL for Win32). Several rules may be applied concurrently.

ANNEX 4. DESCRIPTION OF TECHNIQUES FOR MRF RECOGNITION QUALITY IMPROVEMENT BASED ON THE PENSION FUND OF RUSSIA'S EXAMPLE

One of ways to increase the MRF recognition quality with FineReader 4.0 Rukopis is to use specialized field languages (may be created by user when doing form description) and regular expressions (PERL). The form field description is exemplified below with the form used in the Pension Fund of Russia (the form itself is attached hereto).

The following groups of languages are used in the described form:

- Cities
- Address
- Document Number
- Issuing Authority

as well as individual languages:

- Family Name
- Given Name
- Patronymic Name
- Day of Month
- Month in the Genitive
- Year
- Dictionary Cities
- Non-Dictionary Cities
- Raion
- Oblast
- Country
- Postal Code
- House and Apartment Together
- Building-Apartment
- House Number
- Apartme nt Number
- Russian
- Address Abbreviations
- Telephone
- Document Title
- Document Series
- Other Document Numbers
- City in the Genitive
- Low-Priority City
- Issuing Authority Abbreviations

Family name – control rule: field not empty; language: Family Name (check against an external dictionary of family names; only the symbols of Russian alphabet and "-" are permitted).

Given name – control rule: field not empty; language: Given Name (check against an external dictionary of given names; only the symbols of Russian alphabet and "-" are permitted).

Patronymic name – control rule: field not empty; language: Patronymic Name (check against an external dictionary of patronymic names; only the symbols of Russian alphabet and "-" are permitted).

Date of birth

Day – control rule: field not empty; language: Day of Month (only figures are permitted; regular expressions (PERL) are used).

Month – control rule: field not empty; language: Month in the Genitive (check against an external dictionary of months; only selected symbols of Russian alphabet are permitted).

Year – control rule: field not empty; language: Year (only figures are permitted; regular expressions (PERL) are used).

Place of birth

City (village) – language group: Cities based on the languages Dictionary Cities (check against an external dictionary of cities; only the symbols of Russian alphabet, figures and the symbols ".", "-" and "/" are permitted) Non-Dictionary Cities (only the symbols of Russian alphabet, figures and "-" are permitted; regular expressions (PERL) are used).

Raion – language: Raion (check against an external dictionary of raions; only the symbols of Russian alphabet and "-" are permitted).

Oblast (krai) – language: Oblast (check against an external dictionary of oblasts; only the symbols of Russian alphabet and "-" are permitted).

Country – language: Country (check against an external dictionary of countries; only the symbols of Russian alphabet and "-" are permitted).

Address according to registration

Postal code – language: Postal Code (only figures are permitted; regular expressions (PERL) are used).

Address – language group: Address based on the languages House and Number Together, Postal Code, Building-Apartment, House Number, Apartment Number, Oblast, Raion, Russian, Dictionary Cities, Address Abbreviations and other languages (check against external dictionaries; selected symbols are permitted; certain symbols are prohibited; regular expressions (PERL) are used).

Address of the place of residence (actual)

Postal code – language: Postal Code (only figures are permitted; regular expressions (PERL) are used).

Address – language group: Address based on the languages House and Number Together, Postal Code, Building-Apartment, House Number, Apartment Number, Oblast, Raion, Russian, Dictionary Cities, Address Abbreviations and other languages (check against external dictionaries; selected symbols are permitted; certain symbols are prohibited; regular expressions (PERL) are used).

Telephone – language: Telephone (only figures and "-" are permitted; regular expressions (PERL) are used).

Document

Document title – language: Document Title (check against an external dictionary of document titles; selected symbols of Russian alphabet and "-" are permitted).

Document series – language: Document Series (only figures and selected letters are permitted; regular expressions (PERL) are used).

Document number – language group: Document Number based on the languages Postal Code (only figures are permitted; regular expressions (PERL) are used) and Other Document Numbers (only figures are permitted).

Date of issue

Day – control rule: field not empty; language: Day of Month (only figures are permitted; regular expressions (PERL) are used).

Month – control rule: field not empty; language: Month in the Genitive (check against an external dictionary of months; only selected symbols of Russian alphabet are permitted).

Year – control rule: field not empty; language: Year (only figures are permitted; regular expressions (PERL) are used).

Issuing authority – language group: Issuing Authority based on the languages Cities (a language group), City in the Genitive, Low-Priority City, Oblast, Russian, Issuing Authority Abbreviations, Country and a series of other languages (check against external dictionaries; uncoupled selected symbols are permitted; certain symbols are prohibited; PERL is used).

Date of filling out

Day – control rule: field not empty; language: Day of Month (only figures are permitted; regular expressions (PERL) are used).

Month – control rule: field not empty; language: Month in the Genitive (check against an external dictionary of months; only selected symbols of Russian alphabet are permitted).

Year – control rule: field not empty; language: Year (only figures are permitted; regular expressions (PERL) are used).

Recommendations for checking the recognition accuracy of FineReader 4.0 Handprint Forms

The technique presupposes the following steps in the process of recognition of forms filled out with hand:

1. FineReader 4.0 Handprint Forms installation from CD.
2. Preparation for making a form blank:
   - print several copies of the demo TIFF file from any (e.g. Microsoft Photo Editor from the MS Office suite) viewer or image editor;
   - fill out the blanks so that each letter is in a separate block, the names of months and other words are real, and individual words are separated with a space.
3. Program start (Start\ABBYY FineReader\FineReader 4.0 Handprint Forms or fine32.exe).
4. Opening of the demo package (Forms\Demo\demo.frm) for input of the prepared blanks.
5. Scanning the filled out forms.
   - before scanning, select the following scanner settings:
     - b/w mode (can be called Line Art, OCR, Bi-tonal, etc.);
     - resolution 300 dpi;
     - set brightness level to meet two criteria (1) gray background should be completely eliminated; (2) letter strokes should not be broken:
   - load the filled out blanks into the scanner;

- execute Scan&Read/Scan or Scan&Read/Scan Multiple Pages.

6. Superposition of a template on the entered forms.
7. Recognition of the template-specified blocks into corresponding fields using the command Scan&Read/Recognize All Unrecognized Pages.
8. Accuracy check and correction of the recognized fields.

ANNEX 5. AN INQUIRY TO COGNITIVE TECHNOLOGIES ABOUT A POSSIBILITY TO ADAPT THE SYSTEM FOR REQUIRED FORMS INDEPENDENTLY

INQUIRY

From: Michail Muchnik [SMTP:Muchnik@padco.kiev.ua]
Date: 4 May 1999 a. 16:32
To: michael@cgntv.dol.ru
CC: ayuna@cgntv.dol.ru; support@cgntv.dol.ru
Subject: test ICR Cognitive Form

Dear Sirs,

We were charged to examine the performance of ICR systems. Among these, we also tried to test your CuneiForm. The program was obtained through the courtesy of you Kyiv distributor Mr. Zharkoy. We encountered special difficulties in form creation (*.frm) and block description for recognition (Cognitive FormDesigner). Our attempt to get support in Kyiv has failed. We ask your advice on the matter.

M. Muchnik, S. Mironovich

REPLY

Dear Michail,

The situation with the Cognitive Forms software product is as follows. It is not a box-packed product but a solution that is sold customized. Such customization concerns just the module of your interest, specifically: creation of a standard form blank that you would like to recognize and creation of a description of this form using the Cognitive FormDesigner. Too many nuances should be taken into consideration when creating both blank and description, and since we have a huge experience in this field, it is much easier doing this by ourselves – while giving a 100 percent guarantee of recognition – than training anyone how to do it the right way. That is the reason why the designer is not supplied in the Cognitive Forms software kit: it is an internal product of the company. And it is just that why you have encountered difficulties in the creation of your form description file. Such work is usually done under a contract to supply this software. Every detail would be agreed upon with the customer. In the demo version provided to you by our dealer, there are examples of program operation in the input of payment orders and questionnaires. Should you be interested in our standard form input solution or have any further questions, please feel free to contact me for discussion and solution of these.

Yours sincerely,
Mikhail Potapenko
Sales Manager
Cognitive Technologies Ltd.

phone/fax: (095) 135-5510, 135-8968, 135-5088
e-mail: michael@cgntv.dol.ru
http://www.cognitive.ru

ANNEX 6. ABBYY'S PROPOSALS OF POSSIBLE COOPERATION IN APPLICATION OF THE AUTOMATED TEXT RECOGNITION SYS TEM

To:	Muchnik Mykhailo Manusovich
Company:	PADCO
Phone:	(380-44) 463-76-15, -16, -18
Fax:	(380-44) 463-76-14
Address:	muchnik@padco.kiev.ua
From:	Pakhchanian Aram Bengurovich
Company:	ABBYY / BIT Software
Phone:	+7 095 263-6658
Fax:	+7 095 263-6278
Address:	P.O.B. 19, Moscow 105568 Russia
Subject:	Turnkey installation of an automated data input system

Dear Mykhailo Manusovich,

In order to increase the input speed for the personified information on contributions to the Pension Fund of Ukraine, we offer:

1) to develop machine-readable form blanks for personified information input;
2) to develop a complex for automated input of the Pension Fund of Ukraine's form blanks for further turnkey delivery.

The automated input hardware/software complex will cover the entire process from scanning the paper originals to entering the checked and corrected recognition results into a database. The complex will consist of several locally networked computers with a high-speed scanner connected. The computers will include special software based on the ABBY FineReader Rukopis system for industrial form input.

The quality of the ABBY FineReader technologies has been valued in the Russian and world markets. Our technologies have been licensed and successfully integrated in Siemens Nixdorf (Germany), Primax (the Netherlands), Samsung Electronics (South Korea), NewSoft (the United States) and National News Service (Russia).

Our systems are successfully applied by the Moscow Government, the Central Electoral Committee, the Pension Fund of Russia, the Savings Bank of Russia, the Testing Center attached to the Education Ministry of Russia, the National Registration Company (Norilsk Nickel) and the State Tax Service of Russia. For example, the ratio of correctly recognized documents in the input of a collection of real documents without verification in the Pension Fund of Russia was 99.93 percent.

As far as we know, the FineReader Rukopis system for industrial form input was tested by PADCO and showed a higher efficiency in comparison with a competing system.

Our experience in Russia makes us say quite certainly that application of the ABBYY complex in the Pension Fund of Ukraine will allow a threefold to fivefold increase in the information input speed.

For a trial period – the decision on dates and duration whereof we offer the Pension Fund of Ukraine to take – the automated input complex will be delivered free without any financial liabilities on the side of the Pension Fund.

Yours sincerely,

Aram Pakhchanian

Vice President for corporate projects
ABBYY

Dear Mykhailo Manusovich,

Further to our offer for the Pension Fund of Ukraine that was sent to you earlier, I would like to draw your attention to the following factors that have a special importance for the performance of automated input complex:

1. **The maximum thoroughness and responsibility in preparation of form blanks**. Any minor mistake made at this stage will be multiplied in millions of copies after the blank is approved for use and may turn into an obstacle difficult to overcome in the way to attain a high performance of the complex.
2. **Choice of a good industrial-grade scanner**. Industrial scanners are heavy-duty and rated to withstand continuous operation. The scanning speed is between 40 and 200 pages per minute. The devices of such class automatically process image in real time, thus providing the scanning quality. The feeders of industrial scanners allow loading up to 500 pages. Industrial scanners are manufactured by such companies as Kodak, Bantec, Bell+Howell and Fujitsu. The price of industrial scanners varies between $35,000 and $130,000.

We are willing to render maximum support in solution of both problems.

Yours sincerely,

Andrey Kolpakov
Manager for corporate projects
ABBYY

ANNEX 7. THE RESULTS OF COMPARATIVE TESTS OF THE SYSTEMS FINEREADER BANK (ABBYY) AND COGNITIVEFORMS (COGNITIVE TECHNOLOGIES) IN THE SAVINGS BANK OF RUSSIA

The results of comparative tests of the systems for payment order symbol recognition FineReader Bank and CognitiveForms: payment order input

1. BACKGROUND INFORMATION

Both the technology of CognitiveForms: Payment Order Input and the technology of FineReader Bank are intended for the input and recognition of large numbers of payment documents and creation of a high-quality and high-precision electronic image of payment document. As for their sets of functions, the two systems are practically identical. The document scanning and recognition processes are realized as separate tasks, allowing workplace versatility in case of joint work of users. After the step of document scanning and recognition, the information in the fields of electronic document may be checked and corrected using add-in dictionaries (Bank Identification Code, Correspondent Account, Bank Name, Settlement Account, Taxpayer Identification Number, Client Name, etc.). Erroneous data are highlighted for manual correction by operator. The user interfaces provide for sufficient ease of handling, analyzing and processing the electronic images of documents.

Both systems provide for a possibility to export results in a required format and allow attaching reference databases to make the recognition results more accurate. A number of measures have been implemented to improve the recognition quality in case of any seals or stamps on the document fields and in case of document skew in the process of scanning. Each of the systems provides a special program interface that allows to extend the set of functions and integrate these in bank automated complexes.

2. MATERIALS AND METHODS

The payment order symbol recognition technologies of CognitiveForms and FineReader were tested in the Savings Bank of Russia's Settlement Center on a special bench consisting of:

- computer HP Vectra VL, 200MHz MMX, 32Mb RAM, MS Windows 98;
- scanner ScanJet 4p;
- software for payment documents scanning (NeST) and recognition (CognitiveForms) provided by experts of Cognitive Technologies Ltd.;
- software for payment documents scanning and recognition (FineReader Bank 4.0) provided by experts of ABBYY.

No additional dictionaries were used in testing for check and correction of the results of payment document recognition.

Workers of the Settlement Center offered 50 standard payment documents received from the Central Bank and processed in the Savings Bank. The documents were sorted into two packs:

- Pack 1: 15 A4 documents of a better quality;
- Pack 2: remaining 35 typical documents.

3. TESTING RESULTS

The average scanning time at hand feed and the average recognition time per one document was as shown in the table:

Table 1

Time spent on one document	FineReader Bank	CognitiveForms
Scanning time, s	20	25
Recognition time, s	28 and 50	13.3

Note: The recognition time in the FineReader technology depends on the document quality: the lower recognition time value refers to Pack 1, and the higher value, to Pack 2.

The comparison of the document recognition quality in Pack 1 brought the following results:

Table 2

	Document number	Number of symbols erroneously recognized in document fields		Comments
		FineReader Bank	CognitiveForms	
1	7468	–	1	
2	00376	–	8	xerocopy
3	2455	1	4	small print
4	191	1	11	italic
5	49	–	4	
6	9314	–	1	
7	322	–	1	
8	520	1	5	
9	90	–	4	
10	62	–	4	
11	2213	–	–	
12	482	–	3	
13	1142	–	4	
14	1172	–	4	
15	243	–	4	
	Total:	3	58	

The results of comparing the recognition quality for Pack 2 were as follows:

Table 3

FineReader Bank	CognitiveForms
❑ 9 documents had no erroneous symbols. ❑ 13 documents had 1 to 5 erroneous symbols. ❑ The number of errors in the rest of documents was more than 10, while having 3 to 10 fields recognized completely correct.	❑ 10 documents failed to pass the recognition stage: 1, another document type (memorandum order); 4, insufficient scanning brightness; and 5, significant deviation from the form of payment order. ❑ 6 documents passed the recognition stage but a significant proportion of fields (80%) was not recognized: 3, insufficient scanning brightness; 2, typewriter carbon copies; and 1, narrow font of dot printer. ❑ The number of errors in the rest of documents was more than 10, while having 3 to 10 fields recognized completely correct.

Note: A significant number of the erroneously recognized symbols were related to the presence of stamps and marks on the document fields to be recognized. A deviation from the standard form of payment document, use of fine or narrow fonts, or blurred image considerably increase the error rate. Automatic scanning brightness control should be used to improve the document recognition quality. Attachment of additional dictionaries will significantly decrease the error rate after the recognition stage.

4. GENERAL IMPLICATIONS

The recognition results may be generalized by estimating the *time that operators will need for input of 50 payment order forms (POF)* using:

- FineReader Bank;
- CognitiveForms;
- manual input as presently practiced.

According to the data in Table 2, FineReader Bank made three mistakes in Pack 1, whereas CognitiveForms stumbled 58 times.

In Pack 2, FineReader Bank flawlessly recognized nine POF and made on average three errors in each of other 13, while remaining 13 forms were requiring manual input. In case of CognitiveForms, almost all 35 POF from the second pack required manual input.

The time needed for input with the use of recognition systems is determined by document verification time. The scanning time and the recognition time do not add to the total input time since these processes progress in parallel with verification.

The verification time consists of three components: (T1) the time for reviewing a document to be verified, (T2) the time for correcting each erroneous symbol, and (T3) the time for manual input of the documents not subject to verification. On average, a professional operator has T1˜ 7 s, T2˜ 4 s, and T3˜ 120 s.

Hence, based on the testing results:

1. Input of 50 POF using **FineReader Bank** is estimated as

   37 recognized POF * 7 s +

   + 42 erroneous symbols * 4 s +

   +13 unrecognized POF * 120 s = **33 min**.

2. Input of 50 POF using **CognitiveForms** is estimated as

   15 recognized POF * 7 s +

   + 58 erroneous symbols * 4 s +

   +35 unrecognized POF * 120 s = **75 min**.

3. Input of 50 POF manually would have taken

   50 POF * 120 s = 100 min.

5. CONCLUSION

The tests have shown that the payment document scanning and recognition technologies of FineReader Bank and CognitiveForms are much similar: both of these have the same equipment requirements and have similar document processing functions and operations.

Both of the systems provide for an increase in productivity as compared to manual input. According to the test results, FineReader Bank increases the operator productivity approximately by a factor of three, whereas CognitiveForms: Payment Order Input, by 25 percent.

ANNEX 8. A PLAN FOR IMPLEMENTATION OF THE PROJECT OF TEXT INPUT AND AUTOMATIC RECOGNITION USING SCANNERS (RECOMMENDED FOR DISCUSSION)

1. Development of a scheme of the PCRS interaction and composition, and negotiation on it with PFU.
2. Development, negotiation on and manufacture of a form blank.
3. Purchase of hardware and software.
4. Introduction and testing of hardware and standard software.
5. Development and adjustment of specialized software in terms of integration with AWB-E.
6. Piloting.